A COMPARISON OF GEOCODING BASELAYERS FOR ELECTRONIC MEDICAL RECORD DATA ANALYSIS

Christopher Ray Severns

Submitted to the faculty of the University Graduate School
In partial fulfillment of the requirements
for the degree
Master of Science
in the Department of Geography,
Indiana University

April 2013

Accepted by the Faculty of Indiana University, in partial fulfillment of the requirements for the degree of Master Science.

_____

Jeffrey S. Wilson, Ph.D., Chair

_____

Daniel P. Johnson, Ph.D.,

Master's Thesis
Committee

_____

Pamela A. Martin, Ph.D.,

ACKNOWLEDGEMENTS

I would like to thank the professors and staff of the Geography department for their knowledge and assistance in my pursuit of this degree. Without their expertise and help I would have not been able to accomplish the goal of earning my MS in Geographic Information Science. I would like to especially thank the members of my committee for their input and assistance in completing the research for this paper. I would also like to say a special thanks to Professor Jeff Wilson for his tireless effort and guidance in helping me through this program and with the writing of this paper. Without his help and support I would not have been able to make it through the thesis process.

Perhaps most importantly I would like to thank my wife for her support and encouragement thought-out the process of working towards this degree. Without her helping manage things at home while I was working on school work, this process would have been impossible. I cannot express how grateful I am to her for her support and help. Finally I would like to thank my daughter Piper for waiting until I was almost done with this program to arrive into our lives. Had she shown up much earlier I might not have had the motivation to keep working on the papers and taking classes.

ABSTRACT

Christopher Ray Severns

A COMPARISON OF GEOCODING BASELAYERS FOR

ELECTRONIC MEDICAL RECORD DATA ANALYSIS

Identifying spatial and temporal patterns of disease occurrence by mapping the residential locations of affected people can provide information that informs response by public health practitioners and improves understanding in epidemiological research. A common method of locating patients at the individual level is geocoding residential addresses stored in electronic medical records (EMRs) using address matching procedures in a geographic information system (GIS). While the process of geocoding is becoming more common in public health studies, few researchers take the time to examine the effects of using different address databases on match rate and positional accuracy of the geocoded results. This research examined and compared accuracy and match rate resulting from four commonly-used geocoding databases applied to sample of 59,341 subjects residing in and around Marion County/ Indianapolis, IN. The results are intended to inform researchers on the benefits and downsides to their selection of a database to geocode patient addresses in EMRs.

Jeffery S. Wilson, Ph.D., Chair

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

Identifying spatial and temporal patterns of disease occurrence by mapping the residential locations of affected people can provide information that informs response by public health practitioners and improves understanding in epidemiological research. A common method of locating patients at the individual level is geocoding residential addresses stored in electronic medical records using address matching procedures in a geographic information system (GIS). Geocoding patient addresses creates a model of disease epidemiology that provides estimates of important characteristics such as the overall extent of disease occurrence and locations of disease clusters or hotspots that may not be apparent in databases that lack spatial information. Address matching is now commonly used in health research and the value of this approach is well-noted in the literature. Georeferencing enables visualization of patterns, linking of disease occurrences to potential causal factors, and identifying relationships between clusters of disease and environmental exposures.[1-3] Previous researchers have suggested that advances in GIS technology, analytical methods and availability of high-resolution georeferenced health and environmental data have created unprecedented opportunities to investigate spatial and temporal patterns of disease.[4]

However, as described in a recent request for proposals from the National Institutes of Health (NIH), geocoding can introduce spatial uncertainty in geographic information.[5] Among the important details that need to be considered when utilizing patient address data are geocoding match rate and positional accuracy. Match rate refers to the percentage of total cases that can be associated with a spatial location. Positional accuracy is a measure of the distance between the geocoded location of an object and the actual spatial location of that object.[6] An additional concern is topological accuracy, or whether the spatial relationships of

the geocoded feature are encoded correctly, such as inside the correct census unit. Match rate, positional accuracy, and topological accuracy can be affected by the data set that is used as a basis for geocoding, which is typically a street database. Furthermore, street databases are constantly changing as new roads are added and address information changes (i.e., street names, address ranges and ZIP codes). The best geocoding base layer for a given project can vary depending upon geographic location, spatial scope of the study, and the intended uses of the data. While many studies adopt one particular street database for geocoding, studies that compare variations in results using different geographic base layers are less common.

The purpose of the research presented in this paper is to evaluate and compare the match rate, positional accuracy, and topological accuracy of geocoding results derived using three different street databases and one parcel database. This work is meant to inform future development of geocoding protocols used to process electronic medical record data collected through the Indiana Network for Patient Care (INPC). The INPC is a health information exchange that links electronic medical records from five major hospital systems that includes over 35 hospitals throughout the state, and data from the Indiana State Department of Health and county health departments.[7] The three street databases examined in this project include the Environmental Systems Research Institute (ESRI) StreetMap database, the 2010 TIGER database available through the U.S. Census Bureau, and a street database produced and maintained for the City of Indianapolis by the Indianapolis Department of Metropolitan Development (DMD). In addition to these three street databases, this study also examines parcel-based geocoding using data provided by the Indianapolis DMD. Comparing the geocoding results derived from these different sources can help to identify the advantages and disadvantages associated with each, and inform future implementations of automated geocoding systems designed to feed near-real-time data between healthcare providers and public health practitioners. The address

data used in the study are derived from a sample of 59,345 pediatric patients tested for high blood lead levels (BLLs) between January 1999 and December 2008 at clinics and hospitals located throughout Central Indiana.

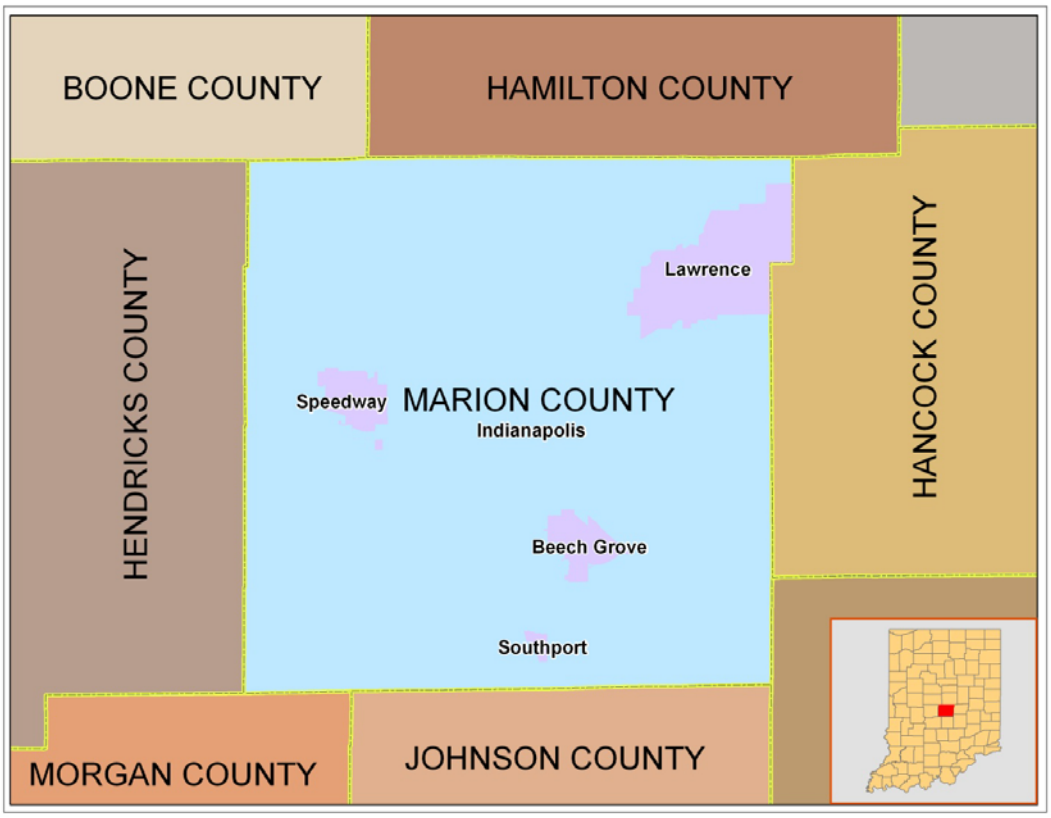

Figure 1: Image of Marion County and surrounding counties. The cities of Indianapolis, Speedway, Beech Grove, Lawrence and Southport are all within the county boundaries.

BACKGROUND

Geocoding has been defined as the process of assigning spatial coordinates to the description of a place by comparing specific elements in a database to those in a geographic reference layer.[8] Though the terms are often used synonymously, address matching is a specific form of geocoding that uses postal address information to estimate the spatial coordinates of a building by using street name, ZIP code, city/town name, and building.[9]

Address matching is now a common process used in epidemiologic research to identify subject locations. However, while it may be common, improved understanding of the effects of spatial uncertainty introduced by geocoding and the subsequent impact on research results are priorities currently emphasized by NIH.[5] Zandbergen discusses several types of errors that can be created in the address matching process.[10] First, positional errors in the street reference data, which is closely related to spatial scale, can lead to propagation of positional error in geocoding results. In other words, if the location of a street network in the GIS is offset from its true location, then geocoding points from this network will be affected by the positional error.

Positional errors can also occur because of representation issues. For example, multilane roads may be represented as a single centerline located along the middle of the network. Positional errors resulting from geocoding based on this simplified representation may be smaller for a two lane roads, but can be larger for multilane roads as shown in Figure 2. Similarly, positional error associated with the representation of the residences can impact analysis. Residence is typically represented as a point location, but the point may not coincide with the actual location where the person lives, such as the location of a specific unit within an apartment complex as seen in Figure 3.

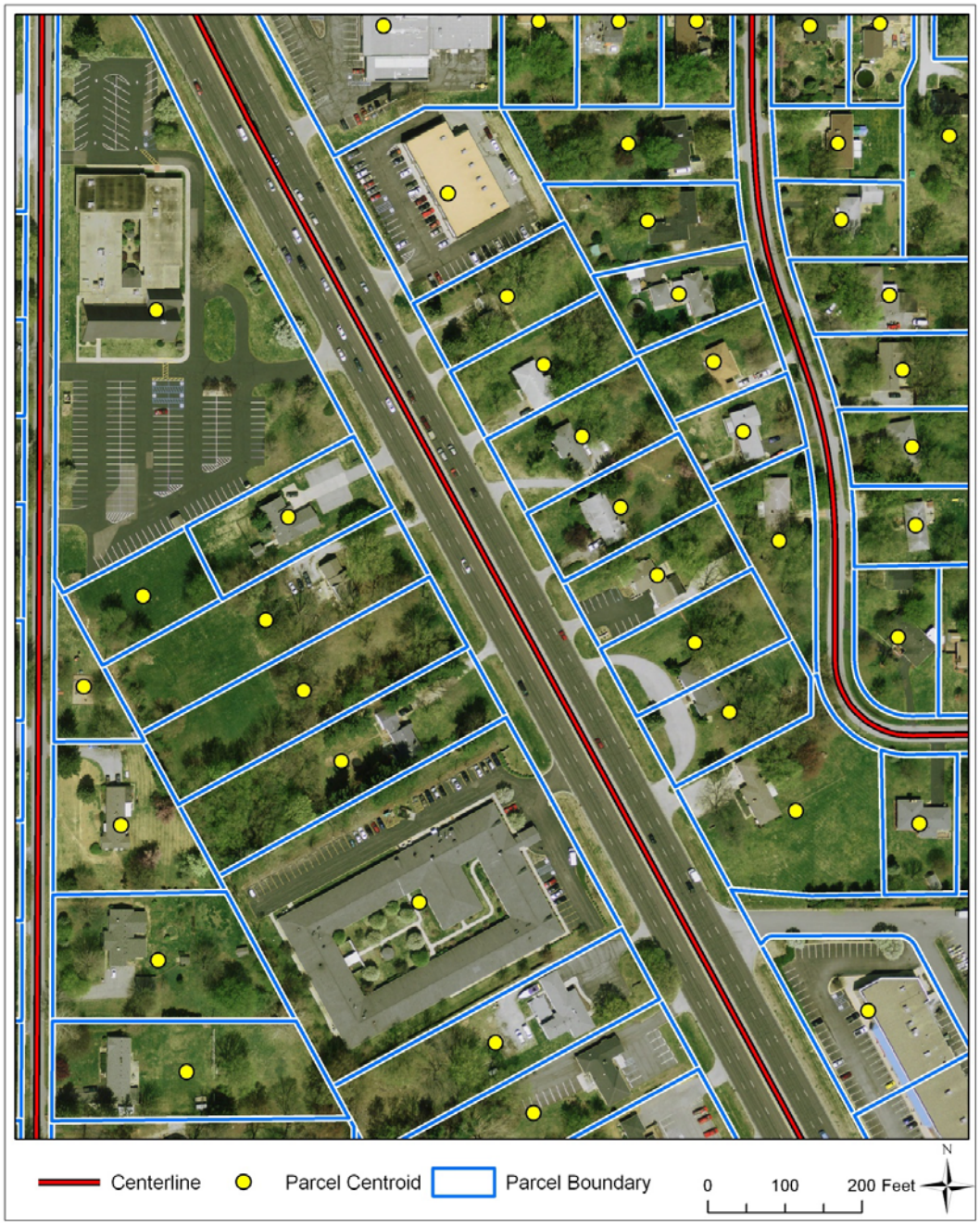

Figure 2: The centerline of a multi-lane road diagonally traverses the figure from upper left to lower right. This image demonstrates how a centerline can have variable distances from a parcel centroid based on the size of the road. Also, an angled road can create parcels that vary in size, thus creating different distances from centerline to parcel centroid.

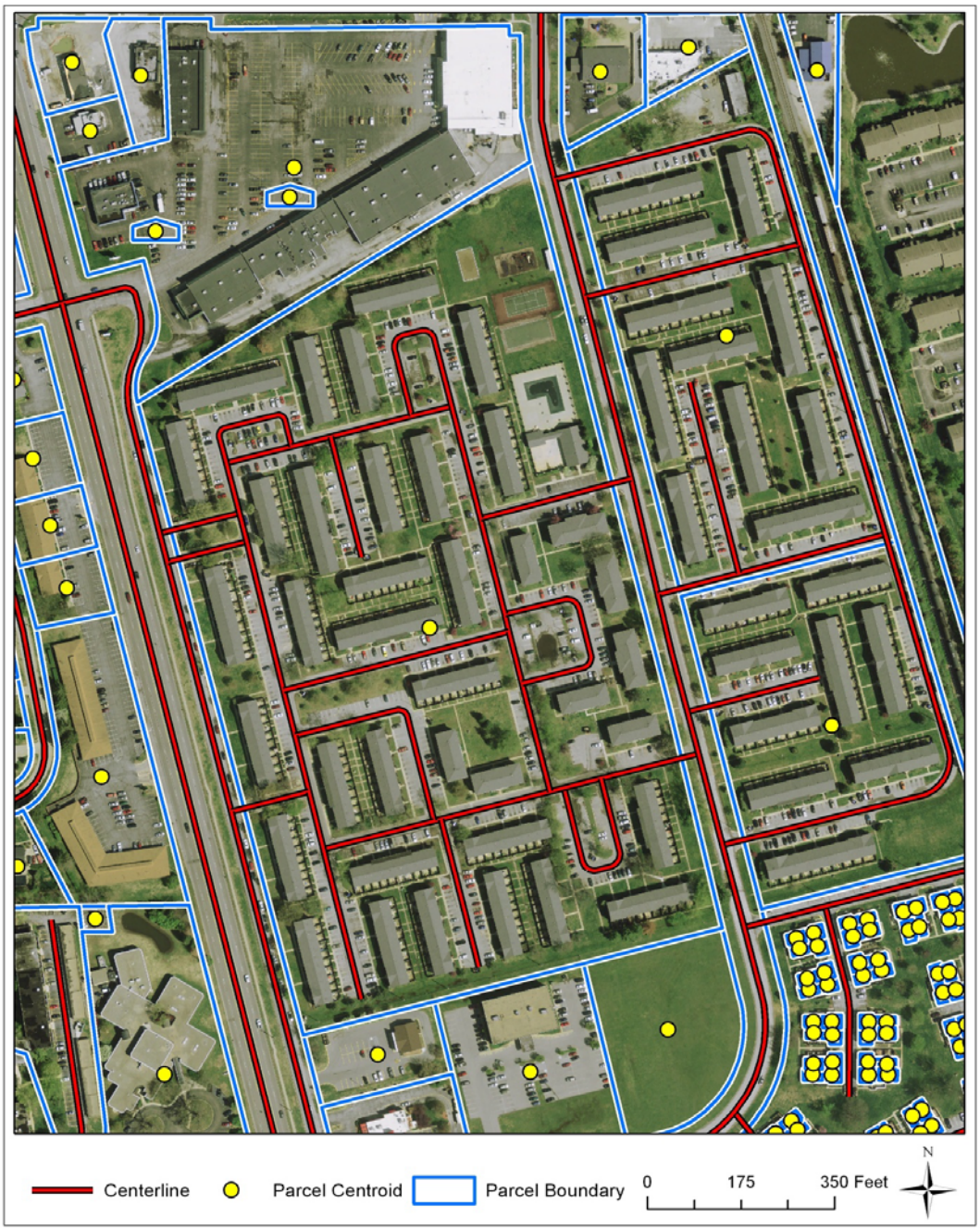

Figure 3: Image provides example of parcel centroids within an apartment complex. This example shows how geocoding to a parcel can produce a very different location than a street centerline as shown in the apartment complex where there is only one parcel but many units within the complex as well as the different centerlines. This can lead to the omission of a geocoded parcel due to there not being an exact match of and address as well as a greater distance of error for a linear geocode to the parcel centroid. These results can influence the measured distance from geocoded point to actual physical location of the dwelling.

As Zandbergen notes, the parcel centroids provide a reliable measure of the location of the residential structure. Based on almost any of the statistics, positional error of parcel

centroids is approximately one order of magnitude smaller than the error of street geocoded locations.[10] Therefore, calculating the difference in location of the parcel centroid location and the street linear geocoded address is an acceptable standard for estimating spatial accuracy of a geocoding baselayer.

Address matching processes typically integrate some form of standardization. Standardization involves parsing the raw addresses into separate elements including the house number, street name, direction prefix and suffix, and street type such as boulevard, lane etc. Standardization organizes raw address data in a consistent format that is congruent with input requirements for most geocoding algorithms and can also be used to identify inconsistencies in data entry such as the use of “AV” vs. “AVE” or “LP” vs. “Loop”. Identifying and correcting for these inconsistencies through standardization enables the GIS to better identify and match the addresses with existing street databases or parcel locators.[11,12] Standardization, however, does not account for missing or incorrect address elements.

Rushton noted the use of a standardized address format leads to fewer errors and describes how the Address Standards Working Group has defined four general types of addresses:[1]

1. Thoroughfare: specifies a location along a linear feature, normally a thoroughfare of some type (e.g., 1225 Rochester Street)

2. Postal: provides a mechanism for mail delivery to a central place without reference to the residence location of an individual (e.g., PO Box 280 Anytown, IA)

3. Landmark: specifies a location through reference to a well known feature (e.g., Madison Square Garden)

4. General: a mix of the first three classes.

The first two classifications are most applicable in the current study, as is in most instances of geocoding health data from electronic medical records, because the majority of patient addresses include street names and building numbers.

Following standardization, the next step in a typical address matching workflow involves comparing the addresses to a reference layer to estimate locations on a map. To complete this process, a geocoding algorithm identifies possible candidates for the location of an address point based on comparison to a reference layer, such as street or parcel data. Most geocoding algorithms provide either a match score or report the number of criteria matched with potential candidate locations. The algorithm utilized by the ESRI's ArcGIS utilizes and alphanumeric index called a soundex. The soundex creates values that are based on specific letters being present in a street name. This soundex determines a match score, which determines whether or not an address meets the predetermined threshold for geocoding to a point.[13] The match scores are not affected by the spelling sensitivity; however, it controls how it considers the spelling results. In a technical paper produced by ESRI they identify the components that make up the scoring criteria. The algorithm considers several factors when determining match score. These include house number, street name, city, pre direction, pre type, suffix direction and suffix type. Each matching component is assigned a point value that, when matched to a baselayer, contributes to the score.[14] The algorithm then compares the address to any location that is at or above the match score threshold set by the user. There are typically three standard levels that are considered for match score results, a score of 100 would be a perfect score, 99 through 80 are generally considered good and less than 80 would not be considered a non-match.[13]

Once a candidate location of suitable quality (a parameter that can be adjusted by the GIS operator) is identified, there are different methods of placing the address point that can

affect the number of addresses matched and the spatial accuracy of the points on the map. The most common method of placement is through liner interpolation along a street segment.

Zandbergen states that the most widely used data model for address matching is street network layer.[4] This model is widely used because it facilitates storing names and address ranges for both sides of a street. Street segments are directionally encoded with a range of addresses (for example 100-199 Main St.). The linear interpolation method of placing points assumes that addresses occur at equal distances from one another along the street segment. For example, 150 Main St. would be assumed to occur near the middle of the segment ranging from 100-199. Additionally, street segments can include even and odd numbers that occur on opposite sides such as dual range addresses. This gives the opportunity to place the point on the correct side of the street. When a spatial offset is used to place the geocoded point, the offset defines the distance perpendicular from the street centerline where the point will be located.

Despite the popularity of the liner interpolation approach, there are issues with accuracy associated with this method. For example, it assumes that all the addresses included within the range for a given street segment actually exist. Additionally, linear interpolation assumes that lots are of equal size and it does not take into account the corner lot dimensions that may be part of intersecting street segments.[15]

Parcel geocoding is another approach to identifying address locations that is examined in this research. Parcel geocoding utilizes property boundaries or centroids that have been attributed with specific addresses. Geocoding using a parcel-based approach involves looking for a match between a parcel address and a patient address in an EMR. If a match is found, a point is located within the boundaries of a given property. Rushton states that in parcel geocoding, a coordinate is normally assigned either to the centroid of the parcel or to the

location of the center of a building footprint on the parcel.[1] Figure four is an example of parcel centroids in an urban neighborhood setting in central Indianapolis. If the address does not match any parcel address, then it is marked as unmatched. Parcel geocoding typically will produce higher spatial accuracy than liner interpolation because it is a more stringent approach that requires a one-to-one match.[1] While parcel geocoding may result in more spatially accurate geocoding, it may also result in lower match rates because of its more strict matching criteria. Miranda et al. suggests that street geocoding often locates general house vicinity but rarely pinpoints the exact housing unit.[16] Parcel geocoding, however, does provide the potential to locate the exact property, providing a more geometrically accurate geocoded location. Figure 4 shows an example of parcel centroids that have been calculated for an urban area located within Marion County.

Figure 4: An example of parcel centroids and boundaries located in an urban area of Marion County. While the centroids are not located exactly on the residential structure, they still provide an accurate representation of the parcel location and a metric useful for accuracy assessment.

In some instances, such as emergency response or identifying subjects exposed to a potential disease-causing agent, spatial accuracy may be of critical importance because the area affected can be discrete. Understanding limitations of geocoding enables researchers to account for positional errors and make corrections prior to data analysis. For example, Zandbergen

examined geocoding positional error and its effect on identifying exposure to traffic-related air pollution among 104,865 children residing in Orange County, Florida.[10] Vehicle emissions are a source of air pollution in all areas, but can be especially high in urban environments and areas proximal to major roads. Zandbergen's study documented that certain pollutants traveled a finite distance from the road system. Therefore, identifying an accurate location of a residence through geocoding was critical to determining whether or not a home would fall within a specified distance of the road network and thus within the pollutants range. Results of this study indicated that median positional error was 41m and that the number of potentially exposed children was consistently overestimated using linear interpolation address matching when compared to parcel-based geocoding.

Rushton compared geocoded locations derived using an address ranging approach from TIGER base layers to the actual locations of residences determined from high-resolution orthoimagery. This study examined approximately 10,000 residences in Carroll County, Iowa. When geocoded locations were compared to the actual locations, the average error was approximately 450m. Rushton concluded that this is significant because geocoded addresses are often used to concentrate a study to a specific area. If the error is too large, it can skew the data and cause relevant data to be excluded from analysis (false negatives) or incorrectly attribute cases that are outside a zone of impact (false positives).[1] Zimmerman et al. found that the largest errors encountered in the geocoding process using TIGER files were attributed to street segments that had correct street names but incorrect address ranges.[2] When considering the use of TIGER for spatial accuracy, it has been suggested that the TIGER system was developed for small scale mapping and is not spatially accurate when high-level spatial accuracy analysis is intended.[17]

In a study using 19,791 addresses, Ratcliffe found that the mean distance between

geocoded points and parcel centroids was 31m in an urban setting using TIGER as a geocoding base layer.[17] He also noted that 5% of the geocoded points were placed in the wrong census unit which creates topological error. Other researchers have reported mean positional errors in addresses geocoded through commercial services between 50m and 300m.[18,19] Whitsel compared four commercial address geocoding services to established longitude and latitude coordinates for residences of participants in the Women's Health Initiative study.[20] The match rates among the vendors ranged from 30% to 98% and average positional errors ranged from 228m to 1,809m. Higher match rates for a given commercial vendor were inversely related to positional accuracy of the point placements.

Other studies also found that geocoding results in urban areas were generally more accurate than in rural areas. This is attributed to shorter street lengths, and more uniform spacing and size of residential parcels within cities.[1,4,18] For example, in Strickland's study conducted in Gwinnett and Fulton Counties in Georgia, it was noted that location error was 35% greater in Gwinnett Co. (predominantly suburban) compared to Fulton Co., which contains a combination of urban and suburban areas, including most of Atlanta's urban core.

DATA AND METHODS

The address data used in this study were derived from a sample of pediatric patients that were tested for elevated blood lead levels. These data were acquired from electronic medical records through the Indiana Network for Patient Care (INPC) based on patient samples collected between January 1999 and December 2008. The sample contains 59,341 listings of patient visits during which blood samples were collected to test for elevated lead levels. A total of 33,631 unique subjects were included in the sample as identified by unique patient identification numbers. A home address was requested with every occurrence of a testing procedure and some patients had multiple listings in the database. Some patients retained the same address throughout the study period, while others moved to different homes in Indiana or out of the state. Demographic characteristics provided in the data include gender, race, and age. The address information associated with each record consists of separate columns for the street, city, state and ZIP.

To prepare the data for analysis, records that had insufficient address information for street-based geocoding were removed (n=11,962). Records were removed if they did not contain street addresses or if only partial address information was provided. For example, if a record only contained a house number and had no street name, or had a street name but no house number, it was excluded. Records that only included a PO Box address were also removed prior to analysis. While PO Boxes are legitimate addresses for mail service, they do not exist within a street or parcel database and therefore are not able to be geocoded. Duplicate addresses that existed within the data were also removed prior to analysis. The rational for removing the duplicate addresses was to avoid misrepresentation of the geocoding match rate. Finally, because the study was focused on comparing geocoding results within the city of Indianapolis derived from local, commercial and federal data sources, records for patients that

did not reside within Marion County, Indiana were excluded. Thus, any address that did not include Indianapolis Speedway, Southport, Lawrence, or Beech Grove (or variations on the spelling of Indianapolis such as Indy, INDPLS, etc.) as the city of residence was excluded. The final analytical sample included 29,301 unique addresses within the county.

Three street databases were evaluated as base layers for address matching in the current study. ESRI StreetMap version 10.0 is a commercial product produced and managed by ESRI that integrates data maintained and updated by NAVTEQ and Tele-Atlas throughout the United States. Frizzelle et al. note that these data are intended to be used for display, routing and geocoding of data in the U.S.[21] The ESRI StreetMap data used in this study were last updated in 2011 according to the associated metadata.[22]

Street data from the 2010 TIGER Line files provided by the U.S. Census Bureau was also evaluated in this study. The TIGER files contain geographic and cartographic information that is intended to assist in the processes of mapping, geocoding and referencing files used in census and survey programs as described on the TIGER website.[23] A benefit to utilizing the TIGER files is that these data are freely accessible from the U.S. Census Bureau.

The third street database evaluated in this study was the centerline street data for the City of Indianapolis, IN. These data are created by the Indianapolis Department of Metropolitan Development (DMD) and depict segments and address ranges within the city and unincorporated areas within the boundaries of Indianapolis and Marion County, including Speedway, Southport, Lawrence and Beech Grove. These data are utilized by the City of Indianapolis for management of the public infrastructure and are updated on a continuing basis.

In addition to the street databases, this study also examined parcel-based geocoding using the Indianapolis/ Marion County, Indiana parcel layer. Similar to the Indianapolis centerline layer, the parcel layer is maintained and updated by the Indianapolis DMD. This layer

includes all parcels within Marion County, including the cities of Beech Grove, Lawrence and Speedway. These data are used in infrastructure and code enforcement applications and are updated on a continual basis. In the current study, parcel-based geocodes served as the most spatially accurate source for identifying residential locations and were used as a basis for comparing the spatial accuracy derived from linear interpolation methods using the three street databases described above.

The 29,301 unique addresses were geocoded using ArcMap 10. Analyst-defined parameters were kept constant using the ArcGIS default values for the three street-based geocoding iterations to facilitate comparison of results: spelling sensitivity was set to 80, minimum candidate score was set to 10, and minimum match score was set to 80 with an offset distance of 10 meters.

Parcel centroids were used as a standard location of reference when estimating the positional accuracy of the points produced by the three street-based geocoding iterations. While the parcel centroid may not be exactly over the housing structure on the property, it does provide a constant standard from which the accuracy measurements can be calculated. These locations are the baseline for which the distances from which the other geocoded points were measured. Measurements of central tendency and dispersion were calculated from these distance measures as indicators of overall positional accuracy. Distance decay graphs were also generated to examine the distribution of errors associated with each base layer.

Unique addresses in Marion County were first geocoded to parcel centroids. This set was used as a reference layer to estimate spatial accuracy of subsequent geocodes derived using the three street database layers. Once the coordinates of the parcel centroids had been calculated, the distances of geocodes from each of the three street database layers were measured using the following formula for straight line distance:

Equation 1.

$$\sqrt{[(x_2 - x_1)^2 + (y_2 - y_1)^2]}$$

These results were then used to determine the average distance from the parcel geocoded location, as well as the median and standard deviation. In determining distances between street-based and parcel-based geocodes, any address geocoded by a street database that did not correspond to an address in the parcel geocoded table could not be included in the calculation. Thus, the removal of the addresses that did not match parcel geocodes resulted in different totals compared for distance calculations associated with each street database. However, independent match rates for all four types of geocoding (one parcel-based and three street-based iterations) were computed.

To determine the topological accuracy, addresses geocoded using the Indianapolis parcel layer were spatially joined to 2000 Census block group polygons which had been topologically snapped to match Indianapolis street centerlines. The spatial join was used to determine which block group each of geocoded addresses resides within. The same process was repeated using U.S. Census block group layers topologically matched to each of the street-based geocoding results. This process created estimates of the topological accuracy of geocoding to the Census block group level using the street databases. After the block groups were determined for each of the three street-based geocoding results, they were compared to the block group IDs resulting from each of the other methods to determine if inconsistencies were observed. Estimates of topological accuracy were computed as the percentage of points placed in the correct census block divided by the total of successfully geocoded points that could be matched to a parcel. The accuracy was compared to the locations of geocoded parcels as this was the most accurate way to ensure that a geocode was in the correct census block.

Figures 5 and 6 are examples of the three different types of geocoded baselayers in an urban and rural/suburban setting. Each set of points is annotated with the range of distances

between the parcel centroid and the street geocoded address to illustrate examples of positional error.

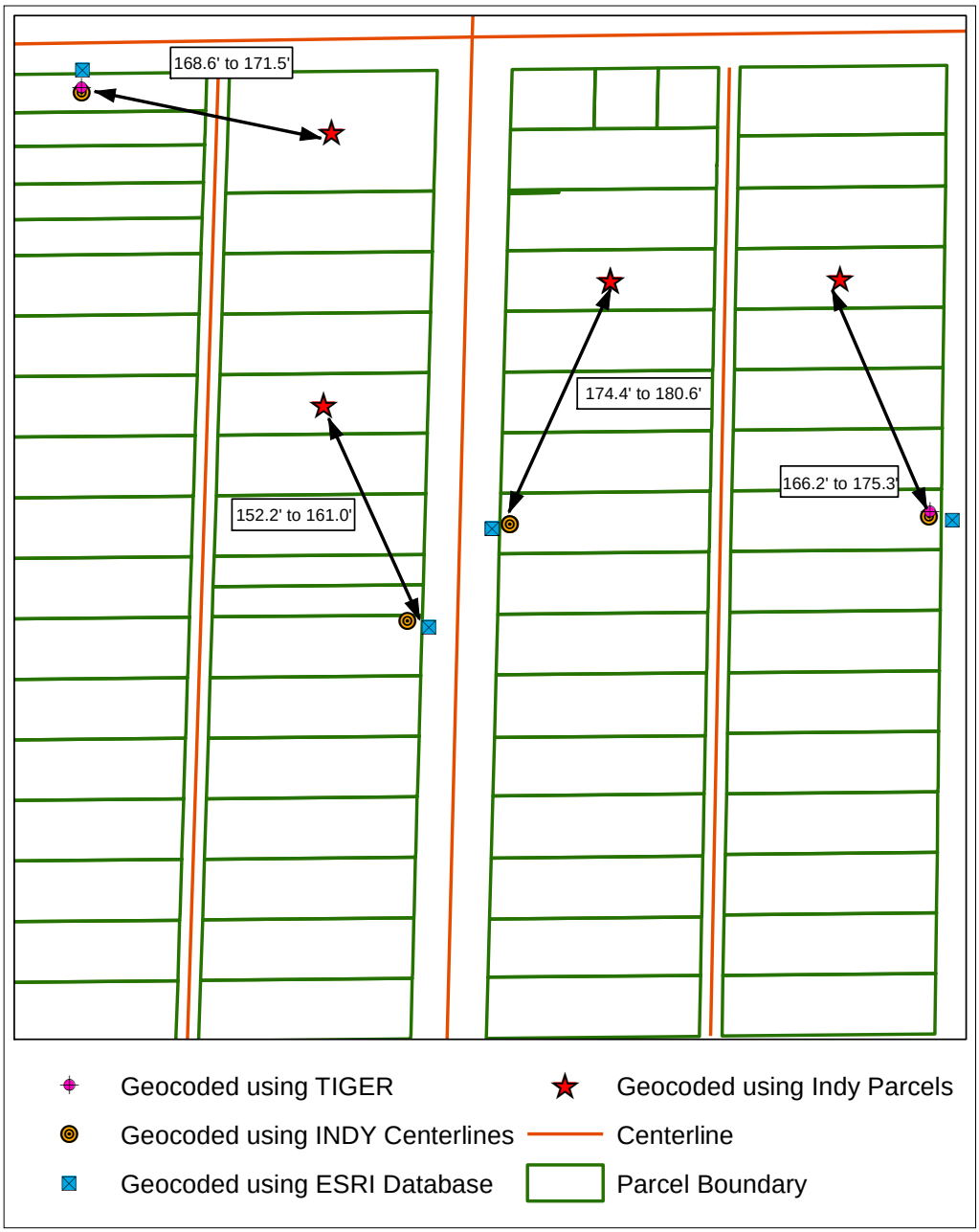

Figure 5: Geocoded addresses in an urban neighborhood in Indianapolis compared to the geocoded parcel centroid addresses. The distances between the centroids and the three street geocoded addresses are given in feet, and the range of distances is given to provide an example of the differences in distances between the methods.

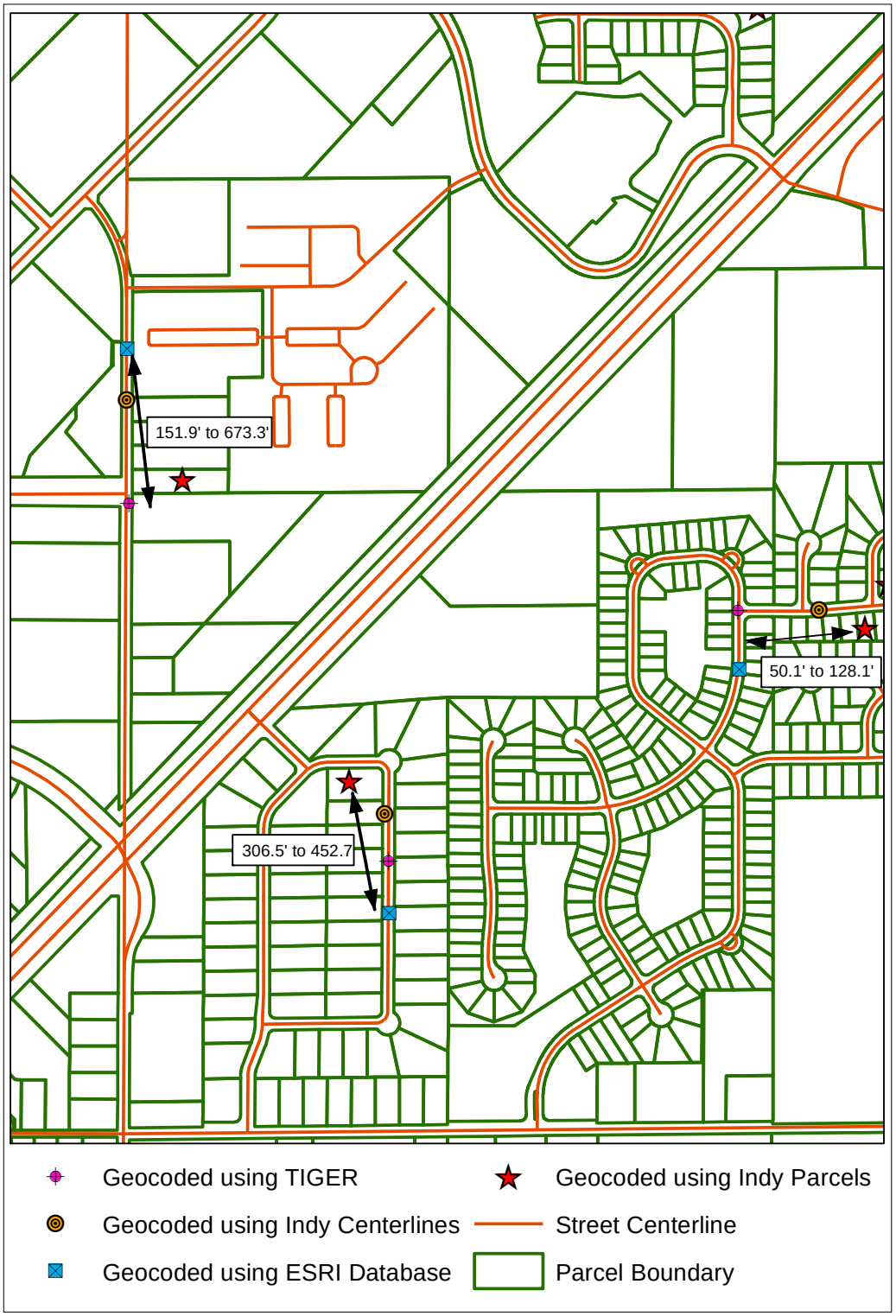

Figure 6: This image is an example of geocoded addresses located in a suburban/ rural area of Indianapolis. The parcel centroid is shown with the distance between the three street geocoded addresses. The distance is given in a range to provide an example of the different results from the three baselayers results from the geocoding process. (The locations of these geocoded addresses have been moved to randomly assigned locations to protect the privacy of the patient data)

RESULTS

Match Rates

Match rates resulting from each of the four geocoding methods tested are summarized in Table 1.

Table 1. Comparison of match rates for the four geocoding base layers.

Comparison of geocoding match rates for utilized base layers. n=29,301				
Baselayer	TIGER	ESRI	Indy DMD	Parcel Centroid
Match Score =100 (%)	2 (>0.01%)	16,847(57.49%)	1(>0.01%)	1,948(6.65%)
Match Score (score 99-80)	18,187(62.06%)	9,264(31.61%)	18,919(64.57%)	9,945(33.94%)
Match Score (Tied 100-80)	147(0.50%)	515(1.75%)	167(0.56%)	957(3.27%)
Unmatched	10,965(37.42%)	2,675(9.12%)	10,214(34.86%)	16,451(56.14%)
Overall Match Rate (%)	62.07%	89.11%	64.57%	40.59%

Spatial Accuracy

Positional accuracy estimates for addresses geocoded using the three street base layers were derived from subsamples of approximately 10,000 address points that were successfully geocoded using the parcel centroids. Table 2 summarizes the minimum, maximum, average, and standard deviation of positional accuracies resulting from geocoding using each of the three street databases by comparing the results to parcel centroid geocodes.

Table 2. Comparison of results of distance calculations from parcel centroids to geocoded addresses.

Summary of distance results for baselayers			
Baselayer	TIGER	ESRI	Indy DMD
Minimum Distance (Feet)	12.03’	20.18’	2.74’
Maximum Distance (Feet)	83,781.12’	83,766.47’	83,794.00
Average Distance (Feet)	365.21’	405.39’	367.23
Standard Deviation (Feet)	2235.14’	2387.77’	2,333.73’

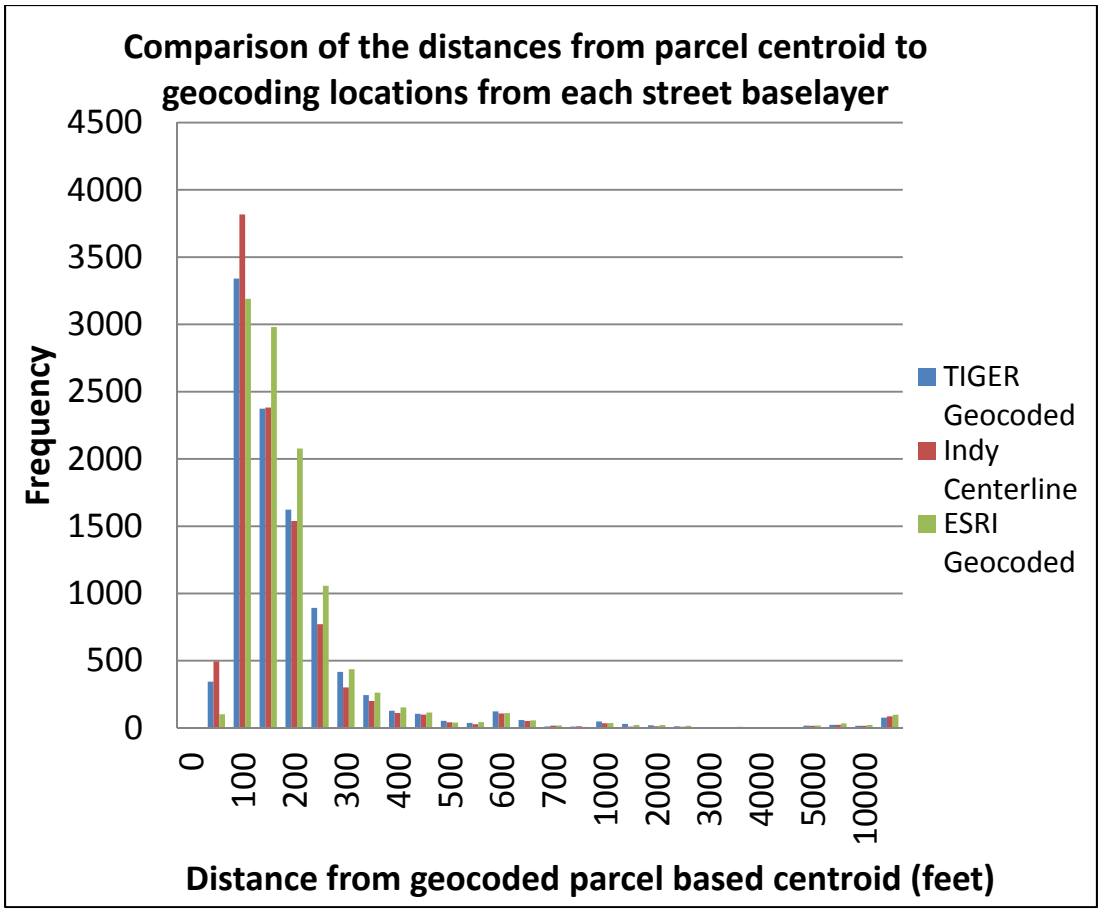

Figure 7: The majority of the distance error occurs in the 0 to 400 foot range which is consistent with expectations based on previous research. There is a small increase in the 600 to 700 foot range as well as a small spike in distances greater than 10,000' which can be attributed to data geocoded to incorrect locations or outside of the county.

Figures 8-10 provide a comparison in the calculated distance error from the parcel centroid to the geocoded locations with the distance of the address from the county center. In examining these two measurements we are able to determine if there is any correlation between distance error in geocoding and distance from the county center. As has been noted, geocoded addresses typically have a higher rate of error in rural areas, it is important to see if that result is apparent in the graphing of these two distances. In the figures below there does not appear to be a strong correlation between the distance error and distance from the county centroid for any of the three baselayers.

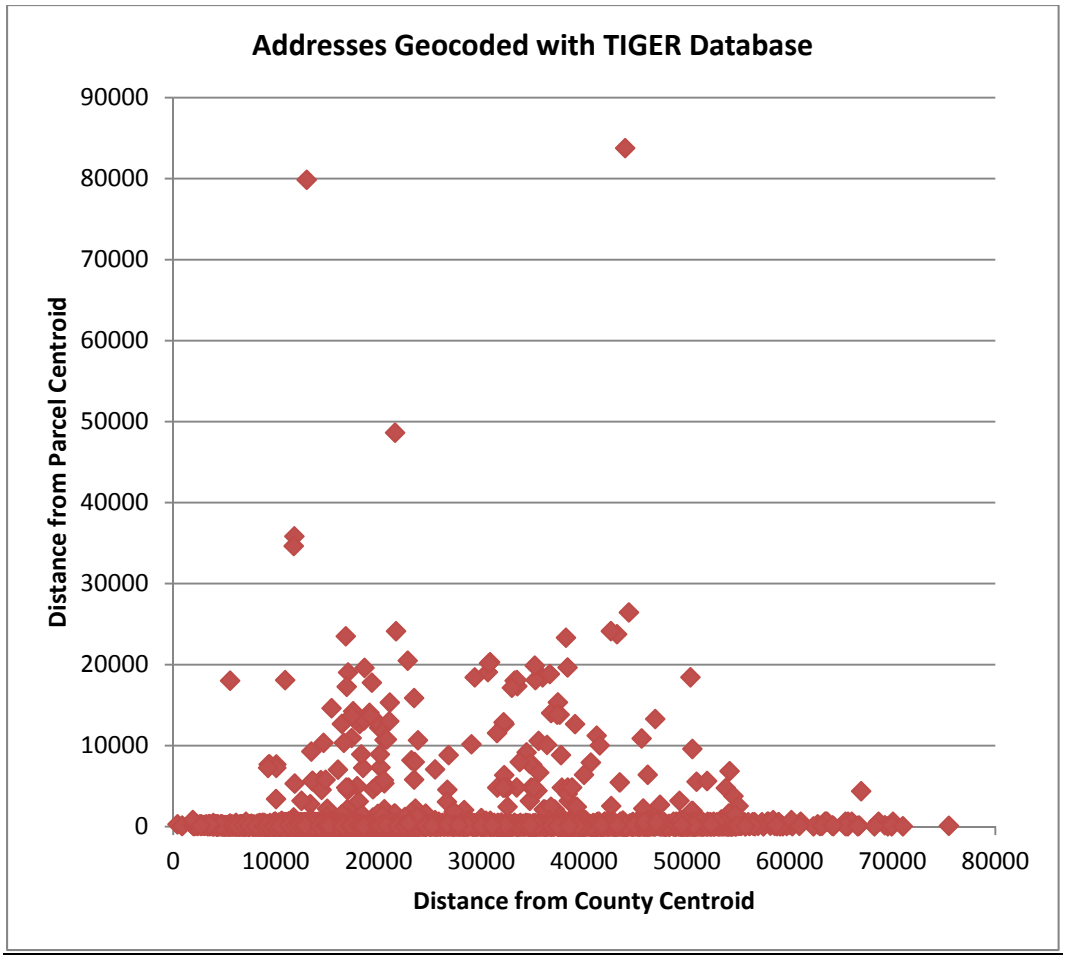

Figure 8: TIGER geocoded address distance error and distance from county centroid.

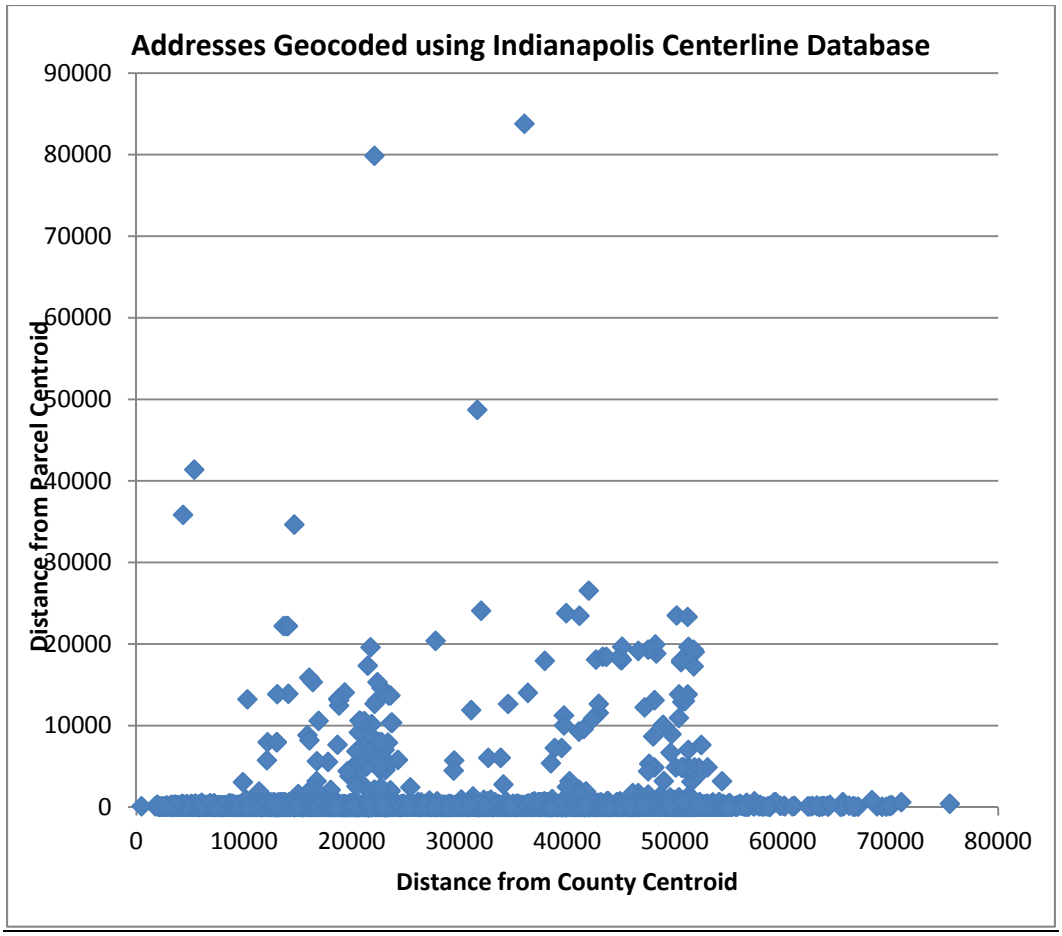

Figure 9: Indianapolis Centerline geocoded address distance error and distance from county centroid.

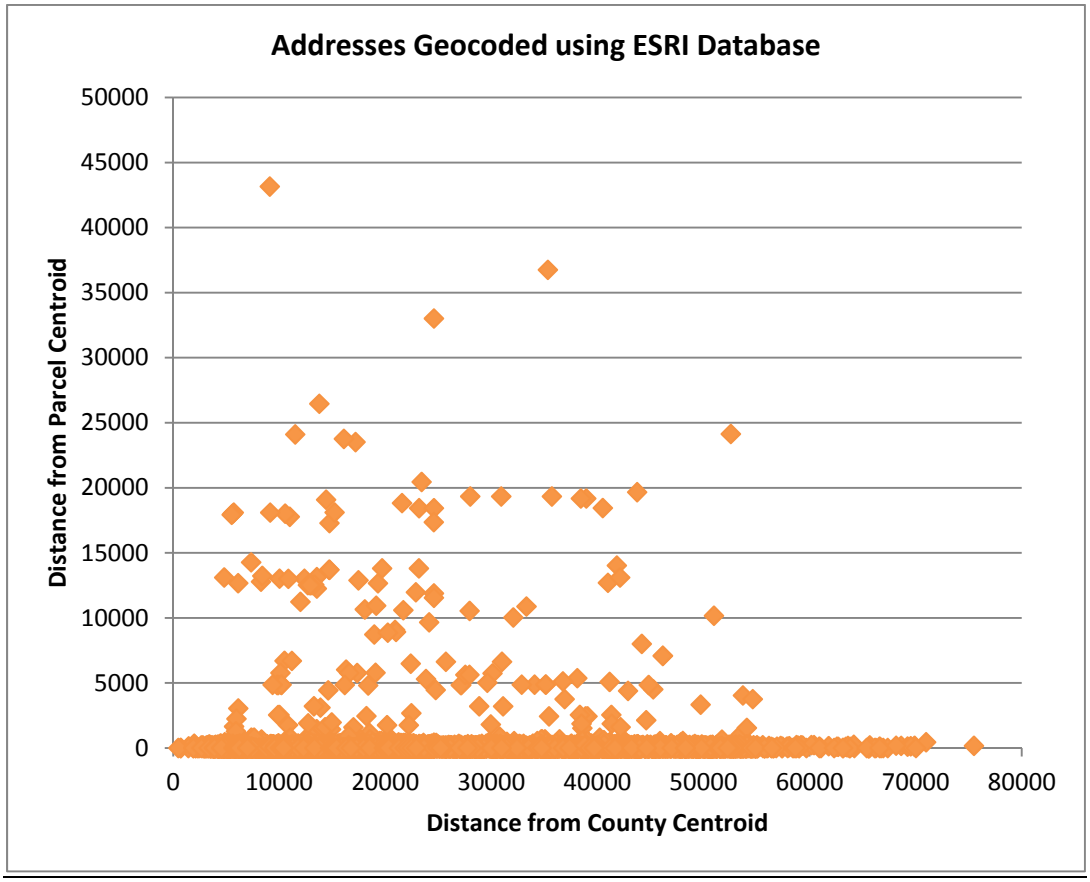

Figure 10: ESRI Address Database geocoded address distance error and distance from county centroid.

Topological Accuracy

Topological comparison of geocoding results derived from the three street-based address matching approaches compared to the parcel centroid method produced the following results. Addresses geocoded utilizing the ESRI Street database placed 10,669 addresses out of 10,944 (97.48%) in the correct block group when compared against the block groups determined by the parcel geocoding and topologically corresponding census geographies. Addresses geocoded utilizing the Indianapolis/ Marion County Street centerline database successfully placed 9,991 addresses out of 10,214 into the correct block group for a total of 97.81%.The addresses

geocoded utilizing the TIGER database successfully geocoded 9,826 addresses out of 10,034 into the correct block group for a match rate of 97.92%.

Table 3. Comparison of topological accuracy at the Census block group level.

Baselayer	TIGER	ESRI	Indy Street
Number of addresses compared	10,034	10,944	10,214
Number in correct block group	9,826	10,669	9,991
Percentage in correct block group	97.92%	97.48%	97.81%

DISCUSSION

Key findings from this study have been summarized in Table 4, which summarizes the geocoding match rates, average positional errors, and topological match rate derived from each of the four base layers tested in this research. The ESRI baselayer produced the highest match rate out of the layers at 89.11%, with the TIGER and Indianapolis Street centerline having similar results with 62.07% and 64.57% respectively. As expected, given the more stringent matching criteria, geocoding using the Indianapolis parcel layer had the lowest match rate at 40.59% which is consistent with previous research.[10] For example, Rushton found that parcel geocoding will produce higher spatial accuracy than liner interpolation but will result in lower match rates due to its stringent approach in requiring a 1 to 1 match.[1]

Table 4. Comparison of geocoding match rates, average position error, and topological match rate for four geocoding base layers.

Summary of results			
Method	Geocoding Match Rate	Average Positional Error (ft)	Topological Match Rate
ESRI	89.11%	405.39	97.48%
TIGER	62.07%	365.21	97.92%
Indianapolis Centerline	64.57%	367.23	97.81
Indianapolis Parcels	40.59%	N/A	N/A

While using the ESRI StreetMap as a basis for geocoding resulted in the highest match rate, it also had the highest average positional error that was about 40 feet greater than the Indianapolis DMD and TIGER centerline base layers. This higher oppositional error may be

attributed to the algorithm utilized by the ESRI StreetMap data. In reviewing the geocoded points from this data there were several that were placed outside of Marion County. Since the ESRI locator is not bound by data only within Marion County it placed some addresses outside of the county. The greater amount of geocoding freedom in the ESRI baselayer to search for addresses contributes to the higher margin of error in the placement. The average positional error resulting from geocoding using the TIGER and Indianapolis DMD databases was nearly equal. By comparing the data as represented in Figure 7, there is a clear trend and spike in the 100' to 200' distances of error which are consistent with the findings of other research. The trend of error decreases as the distance reached the 500' foot error distance as most addresses geocoded with this distance are most likely in rural areas of the county.The Indianapolis Parcel layer was not considered in this comparison because the location of the addresses geocoded using this method were used as a standard to compare the results generated by the other methods.

Topological match rates were very similar among the three street-based layers with less than .50% separating them in terms of the percent of addresses that were matched to the correct block group. The high rate of topological accuracy was not surprising because all three street layers had corresponding Census block group layers that were topologically matched to the centerlines. However, only addresses that could be directly matched to addresses geocoded with the parcel database were tested for topological accuracy, with sample sizes ranging from 10,944 to 10,034.

In reviewing the results from this study, they appear consistent with previous research regarding match rates and spatial accuracy.[10] In Rushton's study he found that the geocoded location in Iowa's Carrol County, a rural section of the state, produced an average positional error of 450m.[1] While Ratcilffe found that average error in geocoding in an urban setting was

31m.[17] Other researchers found errors ranging from 50m to 300m depending on the areas they were geocoding data.Strickland reported positional errors to be 35% higher in rural areas when compared to urban locations.[18-20]

Several limitations of this study should be noted. The research was intended to inform future developments in the geocoding processes used for the Indiana Network for Patient Care (INPC). The INPC is system that has a limited regional focus with the vast majority of patients coming from Central Indiana and Marion County. While the use of the Indianapolis DMD centerline base layer is relevant in the context of geocoding addresses for the INPC, these results cannot be generalized to other locales because the DMD database is unique to Indianapolis/Marion County. However, both the ESRI StreetMap and TIGER base layers are available nationwide and results from this aspect of the study potentially inform broader geocoding accuracy issues.

Similarly, estimates of positional accuracy were limited to subsamples of roughly 10,000 addresses that could be successfully matched to Indianapolis parcel centroids. While some addresses in the southeastern and southwestern portions of the county occurred in areas that are more rural in character, the vast majority of the addresses used in the study were located in suburban and urban neighborhood settings. Thus, results of this study are most relevant to developed areas that tend to produce higher geocoding match rates and positional accuracy according to previous studies published in the related literature.[1,4,8,18,19]

This study was limited to comparing results produced by a single geocoding algorithm implemented in ESRI's ArcGIS software version 10.0. While ArcGIS is popular software, there are numerous other geocoding products available from both open source providers and commercial vendors. Similarly, this study held constant several analyst-defined parameters used in the geocoding process, including spelling sensitivity, minimum candidate score,

minimum match score, and offset distance. In addition to the holding constant the geocoding algorithm and analyst-defined parameters, the current study did not examine the effects of other potential augmentations to the geocoding process, such as the inclusion of alias databases that store alternate names and spellings of street segments.

CONCLUSIONS

Overall the results of this study showed that very comparable rates of topological accuracy were achieved using the three street-based databases. However, these results were limited to a comparison among subsamples of patient addresses that also successfully matched the address of parcel centroids. The match rate, which is arguably a metric of greatest concern to end users of geocoded data, was significantly higher (>25%) when using the ESRI StreetMap database. This result is not surprising given that StreetMap data are derived from NAVTEQ's commercial street databases that are continually updated using a variety of field, image, and database inputs. The market-driven incentives to create value for this commercial product likely contribute to its better performance relative to the two street databases produced by government agencies (TIGER and Indianapolis DMD).

Despite the significantly higher match rate observed when geocoding using the ESRI StreetMap database, this base layer produced an average positional errors exceeding the other two street databases by approximately 40 feet. The ramifications of this observation support a point that was mentioned earlier in this thesis: the specific choices made in a geocoding workflow should be driven by the intended uses of the resulting data. For example, if the intended purpose of geocoding the INPC data examined in this study was to match patient addresses to census or other data sources aggregated to block groups, the relatively high and consistent rate of topological accuracy observed across the street layers combined with significantly higher match rate resulting from the ESRI StreetMap database suggest that this commercial product is a better choice. Conversely, if spatial accuracy was a paramount factor, then the Indianapolis DMD centerline layer may be a better choice, but one that comes at the expense of reduced match rate. Regardless of the base layer that is chosen, users of the geocoding end products should be made aware of these types of tradeoffs through

accompanying data documentation.

While this study did not include data on the demographic characteristics of the patients whose addresses were geocoded, another consideration that end users should consider is the potential for bias in geocoding results. Previous researchers have examined geocoded health data from the central Indiana region and reported that match rates for African American and Hispanic subjects were significantly lower than results for Caucasian / White Non-Latino.[24] Which, according to Sloggett and Joshi could perpetuate racial disparities in health research if not identified.[25]

As there is no perfect system to geocode medical records it is important to know the limitations of each method and consider the effect it can have on future studies. An emerging trend in geocoding medical records is to use of composite geocoding processes. The general idea in this approach is to pass an address through a series of geocoding algorithms, usually with decreasingly stringent spatial criteria. For example, parcel centroids could be the first and most stringent level of geocoding applied, which would likely result in the greater positional accuracy, but at the expense of reduced match rate. Addresses that are not successfully matched to parcels can then be run through a street-based geocoding algorithm. Subsequent iterations could include even more general spatial locators such as the centroid of a town, city, or zip code. These composite approaches inevitably produce higher overall match rates, but end users must be aware of the varying geocoding criteria and integrate these limitations into data analysis considerations.

REFERENCES

1. Rushton, G., *Geocoding Health Data: The use of Geographic Codes in Cancer Prevention and Control, Research and Practice*. 2008, Boca Raton FL: Taylor and Francis Group. 243.
2. Zimmerman, D., et al., *Modeling the probability distribution of positional errors incurred by residential address geocoding.* International Journal of Health Geographics, 2007. **6**(1): p. 16.
3. Jacquez, G. and R. Rommel, *Local indicators of geocoding accuracy (LIGA): theory and application.* International Journal of Health Geographics, 2009. **8**(60): p. 1-17.
4. Zandbergen, P.A., *Influence of geocodin quality on environmental exposure assessment of children living near high traffic roads.* BMC Public Health, 2007. **7**.
5. Health, N.I.O., *Spatial Uncertainty: Data, Modeling, and Communication*. 2011. p. 15.
6. Strickland, M., et al., *Quantifying geocode location error using GIS methods.* Environmental Health, 2007. **6**(10): p. 1-18.
7. McDonald, C., et al., *The Indiana Network for Patient Care: A working Local Health Information Infrastructure.* Health Affairs, 2005. **24**(5): p. 6.
8. Zandbergen, P.A., *A comparasion of address point, parcel and street geocoding tecniques.* Science Direct, 2008. **Computers, Environment and Urban Systems**(32): p. 18.
9. McElroy, J., et al., *Geocoding Addresses from a Large Population-based Study: Lessons Learned.* Epidemiology, 2003. **14**(4): p. 8.
10. Zandbergen, P., *Influence of geocoding quality of environmental exposure assessment of children living near high traffic roads.* BioMed Central Public Health, 2007. **7**(37).
11. Health, W.S.D.O. *Guidelines for Address Matching and Geocoding*. 2007; Available from: http://ww4.doh.wa.gov/gis/geocoding_guideline.htm.
12. Cayo, M. and T. Talbot, *Positional error in automated geocoding of residential addresses.* International Journal of Health Geographics, 2003. **2**(10): p. 1-13.
13. Crosier, S., *Geocoding in ArcGIS*, in *ArgGIS*. 2004, ArcGIS. p. 193.
14. ESRI, *Customizing Locators in ArcGIS 10*. 2010. p. 87.
15. Bakshi, R., C. Knoblock, and S. Thakkar, *Exploiting online sources to accurately geocode addresses.* GIS '04, 2004.
16. Miranda, M., D. Dolinoy, and M.A. Overstreet, *Mapping for prevention: GIS Models for Directing Childhood Lead Poisioning Prevention Programs.* Environmental Health Perspectives, 2002. **110**(9): p. 947-953.
17. Ratcliffe, J., *On the accuracy of TIGER-type geocoded address data in relation to cadastral and census areal units.* International Journal of Geographical Information Science, 2000. **15**(5): p. 473-485.
18. Ward, M., et al., *Positional Accuracy of Two methods of Geocoding.* Epidemiology, 2005. **16**: p. 542-547.
19. Nuckols, J., M. Ward, and L. Jarup, *Using Geographic Information Systems for Exposure Assessment in Environmental Epidemiology Studies.* Environmental Health Perspectives, 2004. **112**(9): p. 1007-1015.
20. Whitsel, E., et al., *Accuracy of commercial geocoding: assessment and implications.* Epidemiologic Perspectives and Innovations, 2006. **3**(8): p. 12.
21. Frizzelle, B., et al., *The importance of accurate road data for spatial applications in public health: customizing a road network.* International Journal of Health Geographics, 2009. **8**(24).

22. ESRI. *North American Address Locator (ArcGIS 10 style)*. 2012; Geocode street addresses in the United States and Canada including street address, ZIP/postal code, city/state/province, and more.]. Available from: http://www.arcgis.com/home/item.html?id=919dd045918c42458f30d2c85d566d68.
23. Census. *TIGER/ Line*. 2010; Available from: http://www.census.gov/geo/www/tiger/.
24. Liu, G., *Differences in self-reported residential location by race, income and education: implications for epidemiologic surveys.* Geocarto International, 2010. **25**(6): p. 429-441.

CURRICULUM VITAE

Christopher Ray Severns

Education
Master of Science, Indiana University-Purdue University Indianapolis
Graduate Certificate, Indiana University-Purdue University Indianapolis
Bachelor of Science, Indiana University
Bachelor of Arts, Indiana University

Research and Training Experience
Geographic Information Systems Analyst Internship, Indianapolis- Marion County Emergency Management Agency, Indianapolis, IN

Professional Experience
INDIANAPOLIS DIVISION OF HOMELAND SECURITY, Indianapolis, IN (November 2010-Present)
GIS Program Director (November 2010- Present)

INDIANA UNIVERSITY PURDUE UNIVERSITY INDIANAPOLIS, Indianapolis, IN (October 2007 – November 2010)
Training and Documentation Specialist (August 2010 – November 2010)
Financial Aid Advisor (October 2007 – August 2010)

INDIAN CREEK MIDDLE SCHOOL, Trafalgar, IN (November 2005 – August 2006)
7th Grade Social Studies Teacher (November 2005 – August 2006)

PARKVIEW MIDDLE SCHOOL, Jeffersonville, IN (November 2004 – June 2005)
7th Grade Social Studies Teacher